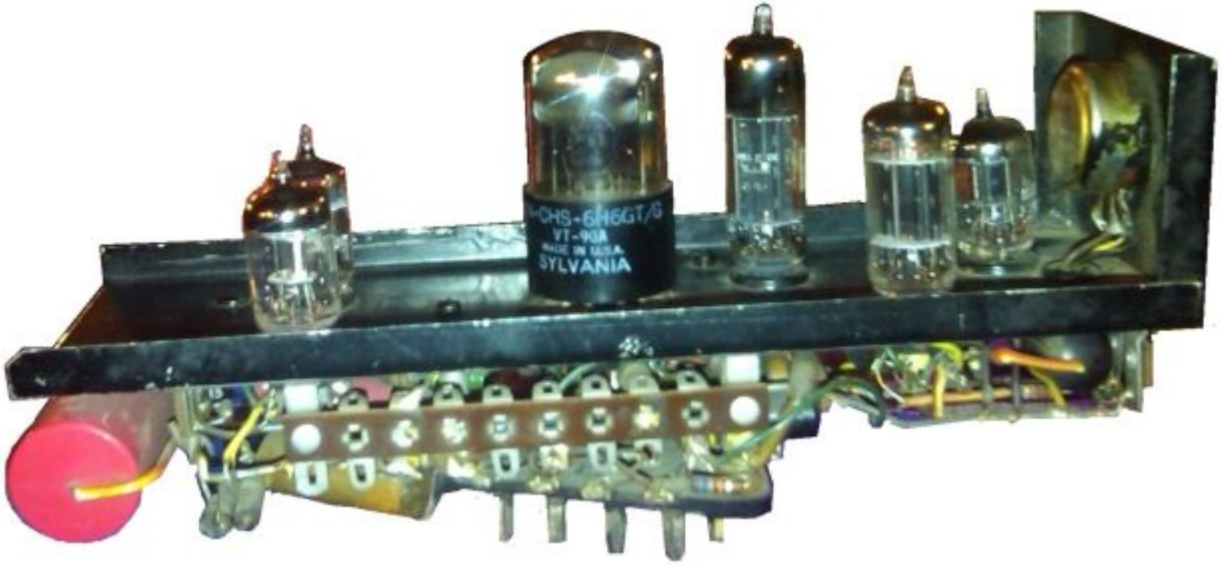
Kubeflow

Towards Kubeflow 1.0: Bringing a Cloud Native Platform for ML to Kubernetes

2019/05/22

Jeremy Lewi (jlewi@google.com)
David Aronchick(daaronch@microsoft.com)



SNARC Maze Solver
Minsky / Edmonds
(1951)

2000

2006

NumPy

2007

NumPy

theano

2008

NumPy

theano

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

2010

NumPy

theano

scikit learn

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

2013

NumPy

theano

scikit learn

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

Caffe

2014

NumPy

DL4J

theano

scikit learn

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

Caffe

2015

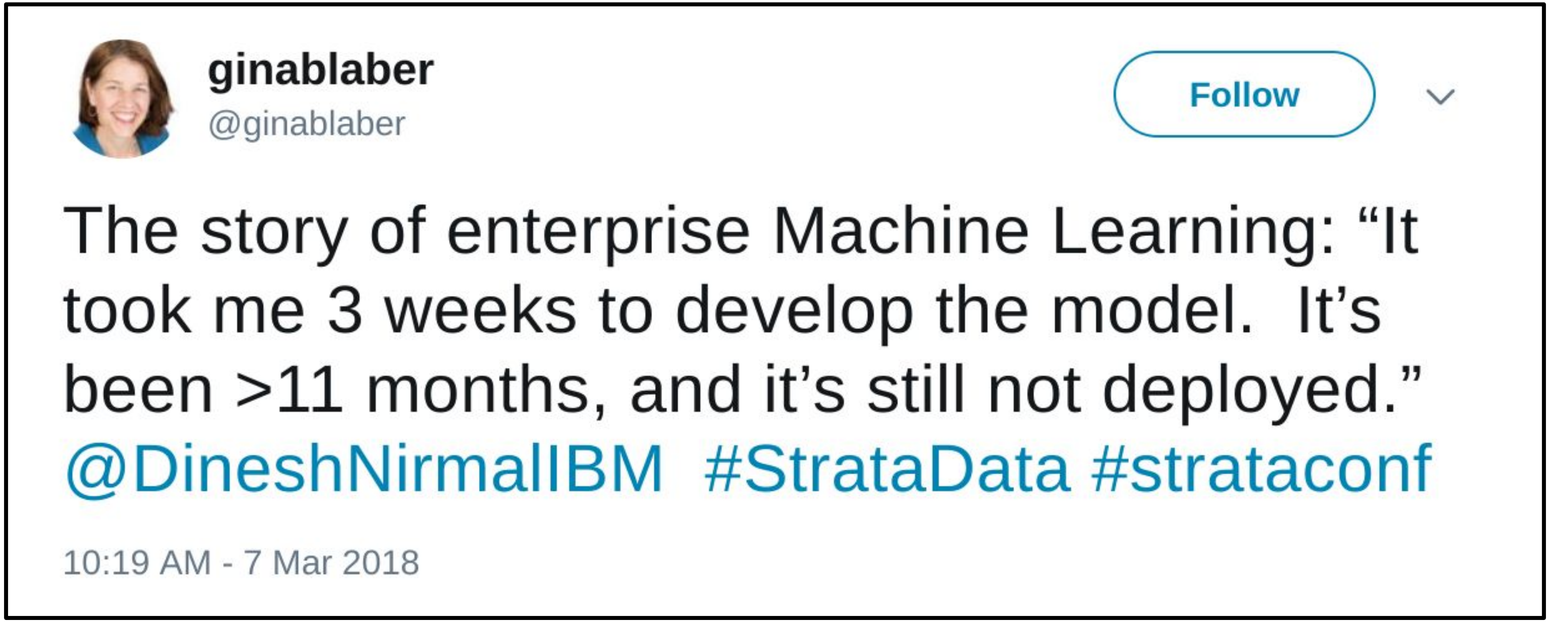
NumPy
DL4J
theano
PyTorch
Keras
Chainer
scikit learn
APACHE mxnet™
pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$
Microsoft CNTK
Caffe

Today

Caffe2
NumPy
DL4J
theano
PyTorch
Chainer
Keras
scikit learn
MATLAB
pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$
APACHE mxnet™
Microsoft CNTK
Caffe
R

One More ML Solution

One More ML Solution???



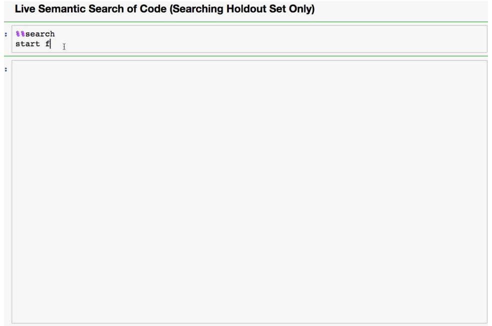
ginablaber
@ginablaber

The story of enterprise Machine Learning: “It took me 3 weeks to develop the model. It’s been >11 months, and it’s still not deployed.” @DineshNirmalIBM #StrataData #strataconf

10:19 AM - 7 Mar 2018

GitHub Natural Language Search

Prototype MVP With Demo In Jupyter Notebook: **2 Weeks**

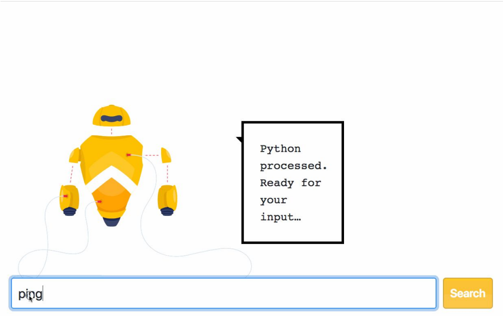


Live Semantic Search of Code (Searching Holdout Set Only)

```
%%search
start f
```

https://github.com/hamelsmu/code_search

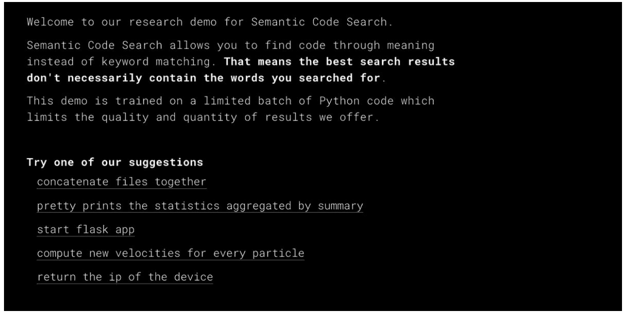
Demo with front-end mockup with blog post: **+3 Days**



Python processed. Ready for your input…

https://towardsdatascience.com/semantic-code-search-3cd6d244a39c

Experiments.Github.Com: **+3 Months**



Welcome to our research demo for Semantic Code Search.

Semantic Code Search allows you to find code through meaning instead of keyword matching. **That means the best search results don't necessarily contain the words you searched for.**

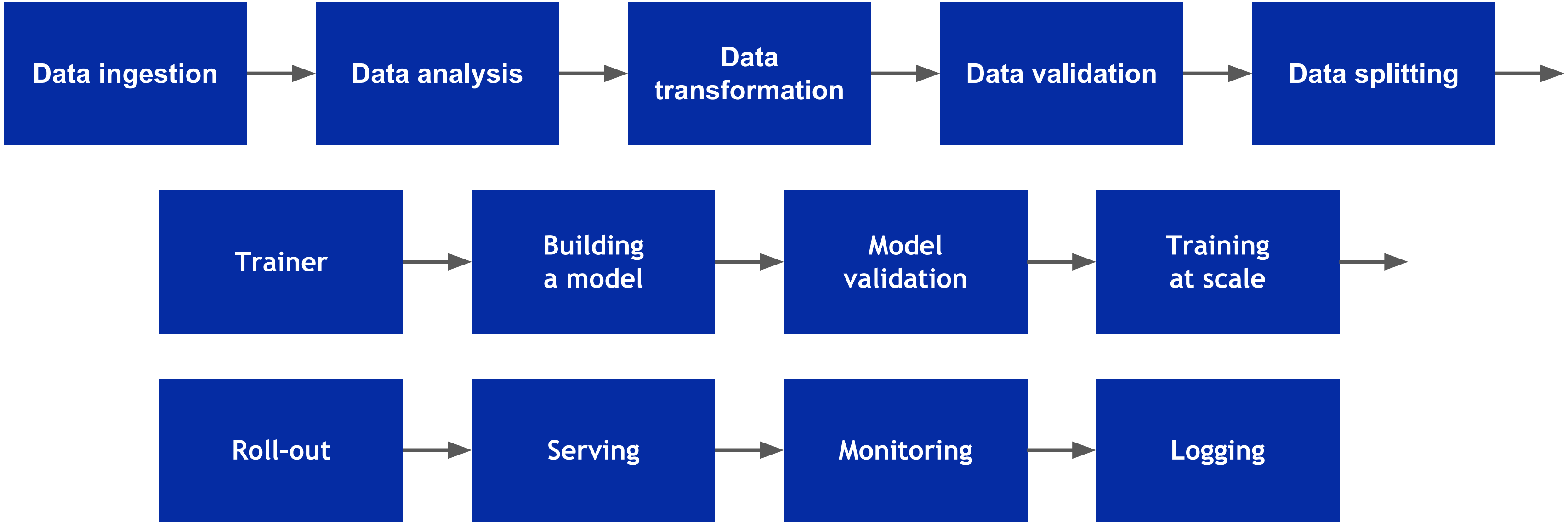
This demo is trained on a limited batch of Python code which limits the quality and quantity of results we offer.

Try one of our suggestions

- concatenate files together
- pretty prints the statistics aggregated by summary
- start flask app
- compute new velocities for every particle
- return the ip of the device

https://experiments.github.com/

Building a model



Data ingestion → Data analysis → Data transformation → Data validation → Data splitting →

Trainer → Building a model → Model validation → Training at scale →

Roll-out → Serving → Monitoring → Logging

Four Years Ago…

Google and Containers

Everything at Google runs in a container.

Internal usage:

- Resource isolation and predictability
- Quality of Services
    - batch vs. latency sensitive serving
- Overcommitment (not for GCE)
- Resource Accounting

We start over 2 billion containers per week.

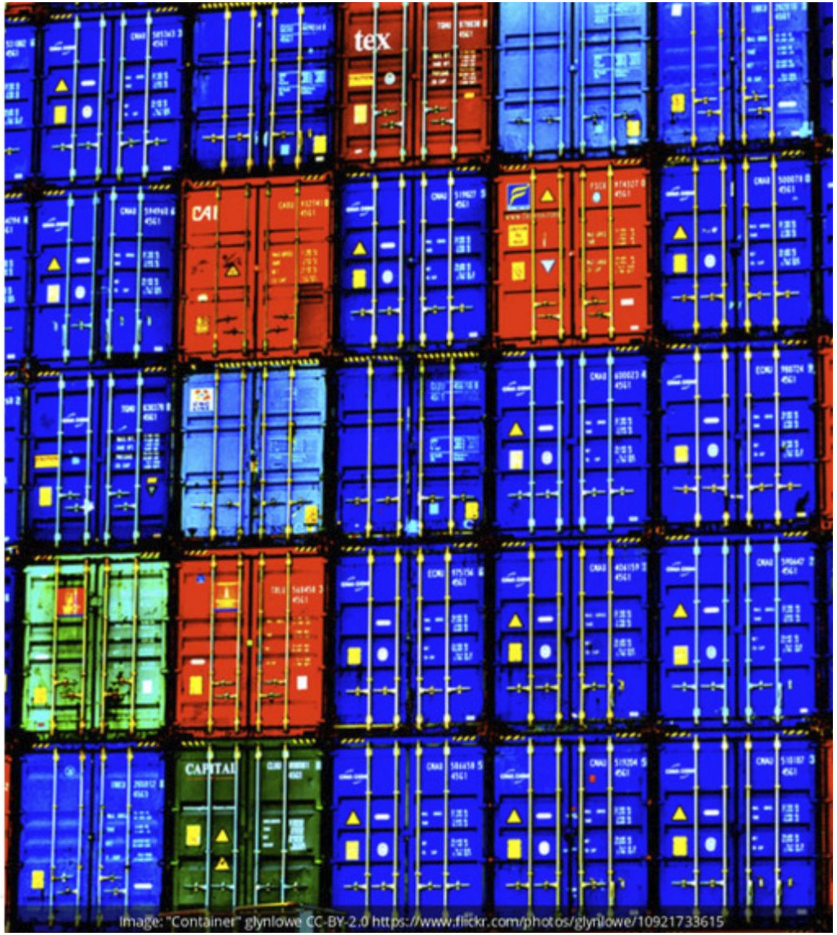


Image: "Container" glynlowe CC-BY-2.0 https://www.flickr.com/photos/glynlowe/10921733615

Kubernetes

Cloud Native Apps

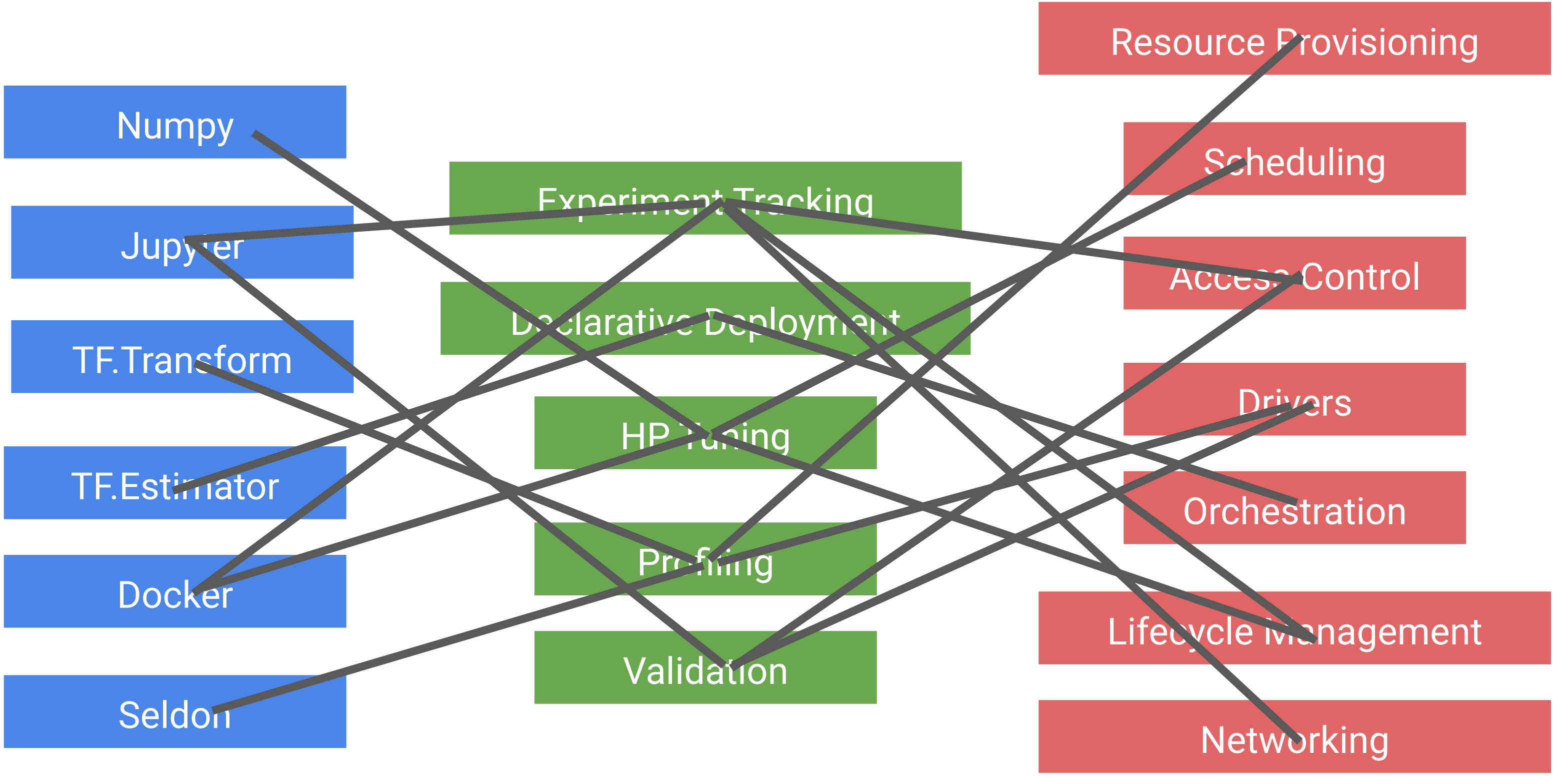
Can we use Kubernetes to fix this?

Oh, you want to use ML on K8s?

First, can you become an expert in ...

- Containers
- Packaging
- Kubernetes service endpoints
- Persistent volumes
- Scaling
- Immutable deployments
- GPUs, Drivers & the GPL
- Cloud APIs
- DevOps
- ...

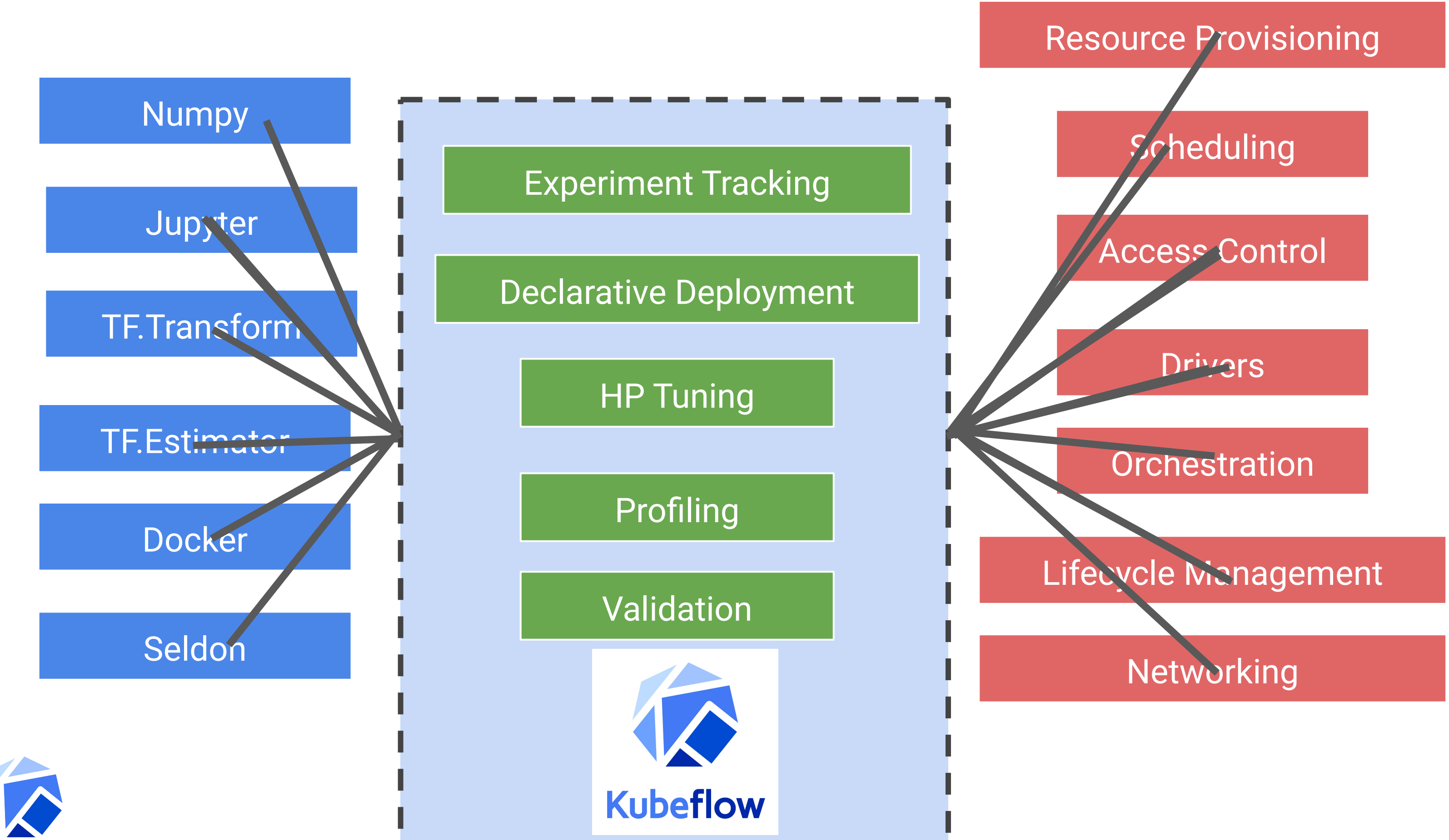
Cloud Native ML?


Numpy
Jupyter
TF.Transform
TF.Estimator
Docker
Seldon
Experiment Tracking
Declarative Deployment
HP Tuning
Profiling
Validation
Resource Provisioning
Scheduling
Access Control
Drivers
Orchestration
Lifecycle Management
Networking

Kubecon 2017

Introducing Kubeflow

Make it Easy for Everyone to Develop, Deploy and Manage Portable, Distributed ML on Kubernetes


Numpy
Jupyter
TF.Transform
TF.Estimator
Docker
Seldon
Experiment Tracking
Declarative Deployment
HP Tuning
Profiling
Validation
Kubeflow
Resource Provisioning
Scheduling
Access Control
Drivers
Orchestration
Lifecycle Management
Networking

Timeline

Dec
2017

Introduce Kubeflow
JupyterHub
TFJob
TFServing

May 2018

Kubeflow 0.1
Argo
Ambassador
Selldon

Aug
2018

Kubeflow 0.2
Katib -HP Tuning
Kubebench
PyTorchs

Sep 2018

Contributor
Summit

Oct 2018

Kubeflow 0.3
kfctl.sh
TFJob v1alpha2

2019
Jan

Kubeflow 0.4
Pipelines
JupyterHub UI refresh
TFJob, PyTorch beta

2019 April

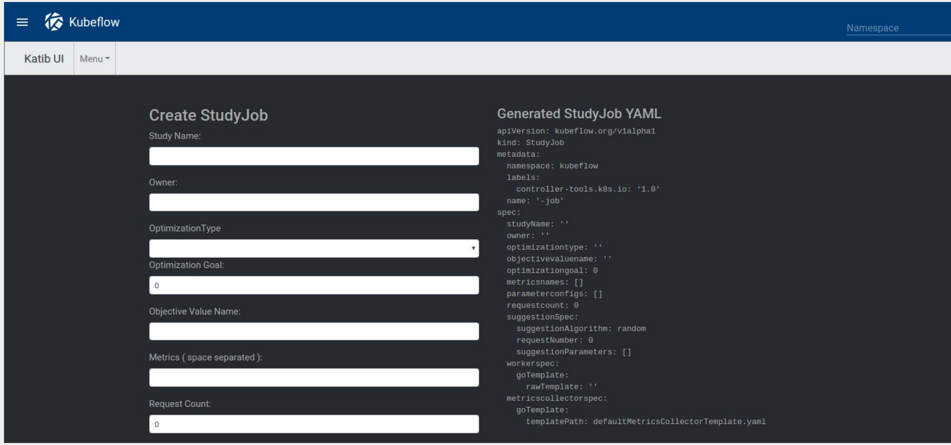
Kubeflow 0.5
Fairing
Jupyter WebApp + CR

Kubeflow is a Cloud Native Platform for ML

Tenets

- **Composable -** Use the libraries/frameworks of your choice
- **Scalable -** number of users & workload size
- **Portable -** on prem, public cloud, local

Hyperparameter Tuning



Create StudyJob

Study Name:

Owner:

OptimizationType

Optimization Goal: 0

Objective Value Name:

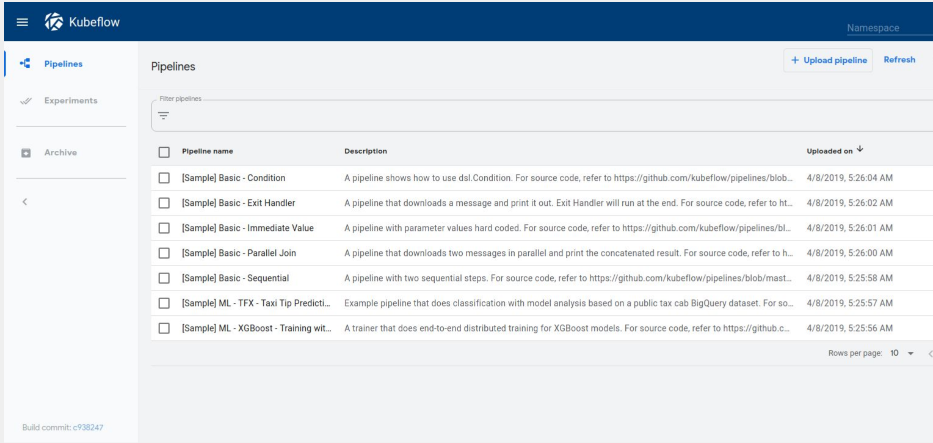
Metrics (space separated):

Request Count: 0

Generated StudyJob YAML

```yaml
apiVersion: kubeflow.org/v1alpha1
kind: StudyJob
metadata:
  namespace: kubeflow
  labels:
    controller-tools.k8s.io: '1.0'
  name: '-job'
spec:
  studyName: ''
  owner: ''
  optimizationtype: ''
  objectivevaluename: ''
  optimizationgoal: 0
  metricsnames: []
  parameterconfigs: []
  requestcount: 0
  suggestionSpec:
    suggestionAlgorithm: random
    requestNumber: 0
    suggestionParameters: []
  workerspec:
    goTemplate:
      rawTemplate: ''
  metricscollectorspec:
    goTemplate:
      templatePath: defaultMetricsCollectorTemplate.yaml
```

Pipelines



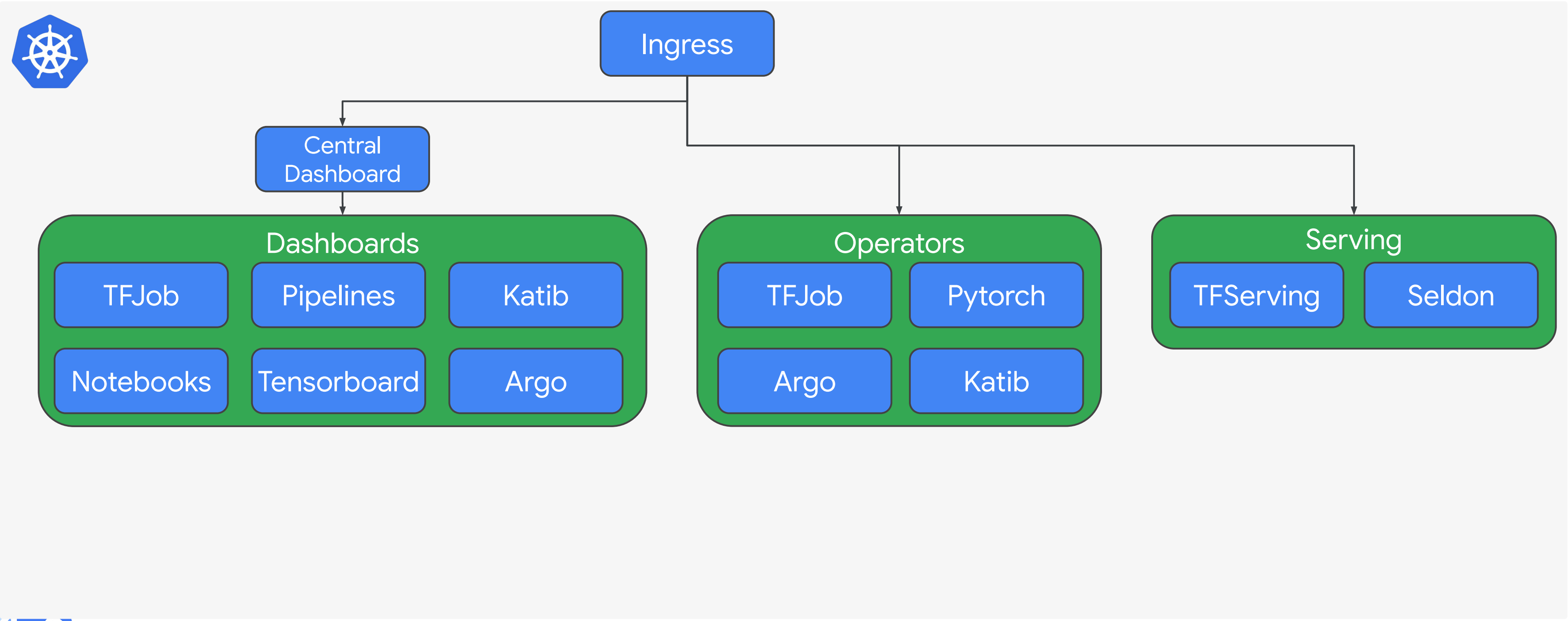
Pipelines

□	Pipeline name	Description	Uploaded on ↓
□	[Sample] Basic - Condition	A pipeline shows how to use dsl.Condition. For source code, refer to https://github.com/kubeflow/pipelines/blob...	4/8/2019, 5:26:04 AM
□	[Sample] Basic - Exit Handler	A pipeline that downloads a message and print it out. Exit Handler will run at the end. For source code, refer to ht...	4/8/2019, 5:26:02 AM
□	[Sample] Basic - Immediate Value	A pipeline with parameter values hard coded. For source code, refer to https://github.com/kubeflow/pipelines/bl...	4/8/2019, 5:26:01 AM
□	[Sample] Basic - Parallel Join	A pipeline that downloads two messages in parallel and print the concatenated result. For source code, refer to h...	4/8/2019, 5:26:00 AM
□	[Sample] Basic - Sequential	A pipeline with two sequential steps. For source code, refer to https://github.com/kubeflow/pipelines/blob/mast...	4/8/2019, 5:25:58 AM
□	[Sample] ML - TFX - Taxi Tip Predicti...	Example pipeline that does classification with model analysis based on a public tax cab BigQuery dataset. For so...	4/8/2019, 5:25:57 AM
□	[Sample] ML - XGBoost - Training wit...	A trainer that does end-to-end distributed training for XGBoost models. For source code, refer to https://github.c...	4/8/2019, 5:25:56 AM

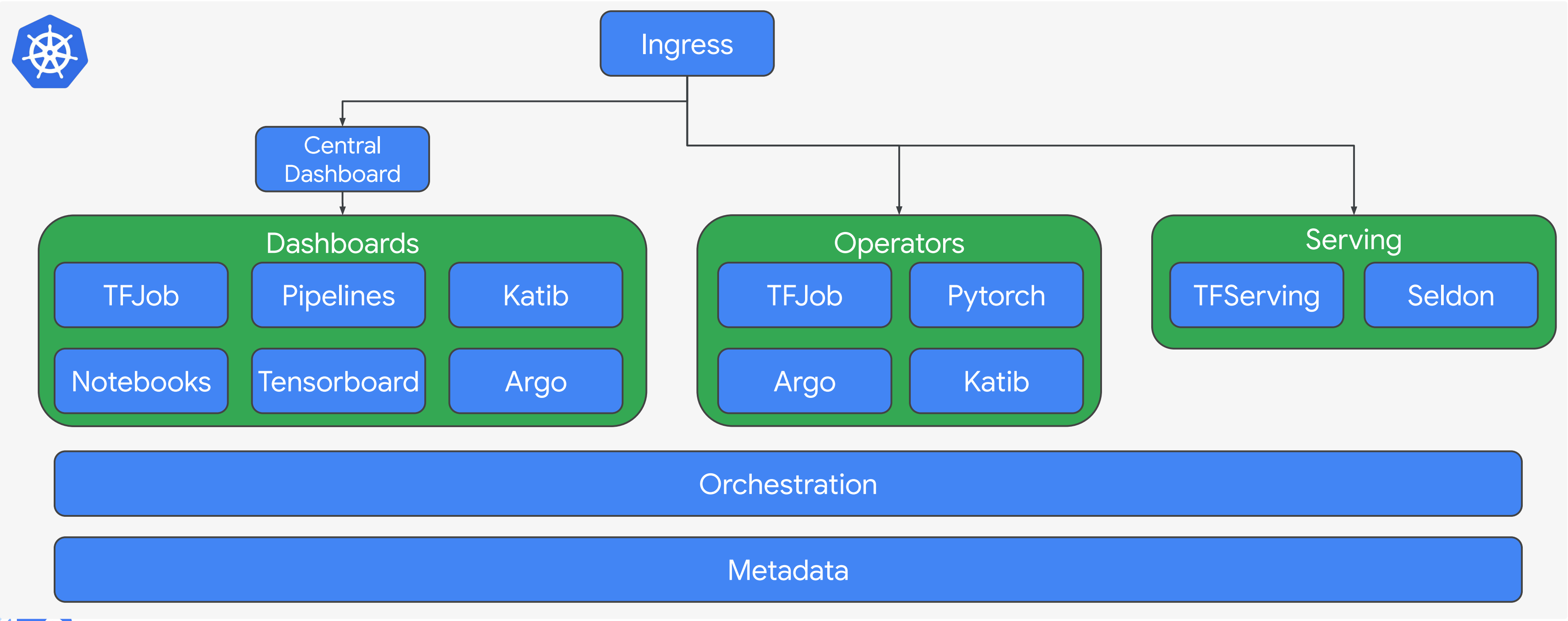
Rows per page: 10

Build commit: c938247

Kubeflow Architecture



Kubeflow Architecture



Momentum!

Arrikto Alibaba Cloud 蚂蚁金服 ANT FINANCIAL argo Caicloud

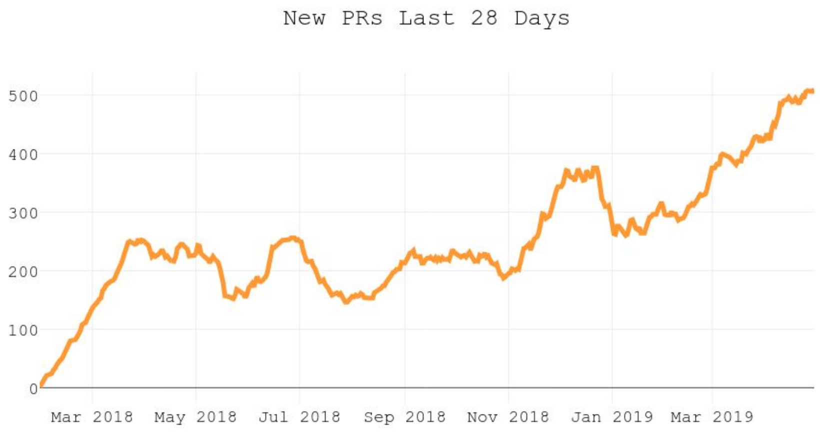
CANONICAL cisco DATAWIRE DELL GitHub GO-JEK

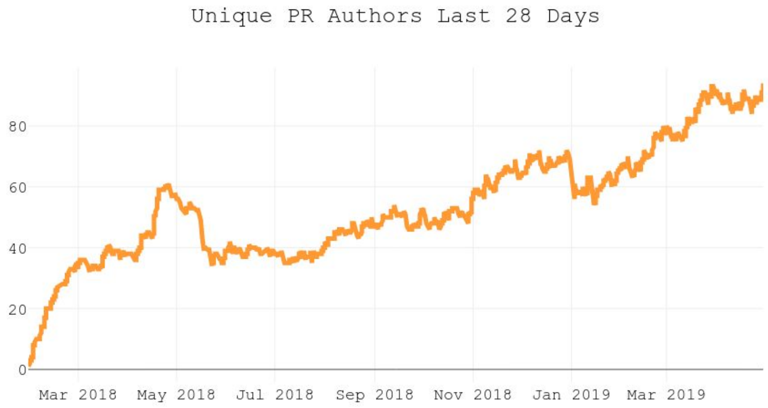
Google heptio HUAWEI IBM iguazio intel

Microsoft MOMENTA One Convergence Pachyderm PRIMER

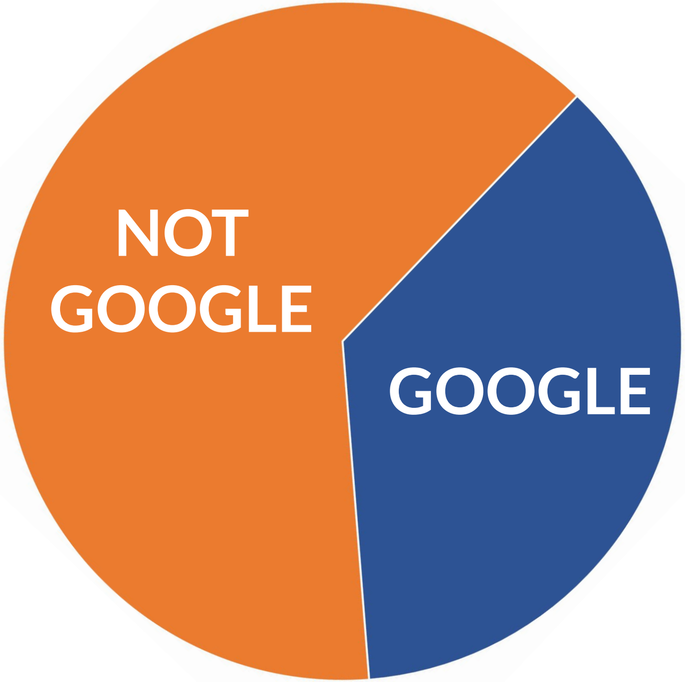
jupyter redhat SELDON shopify tu simple

UBER 云知声 Unisound weaveworks

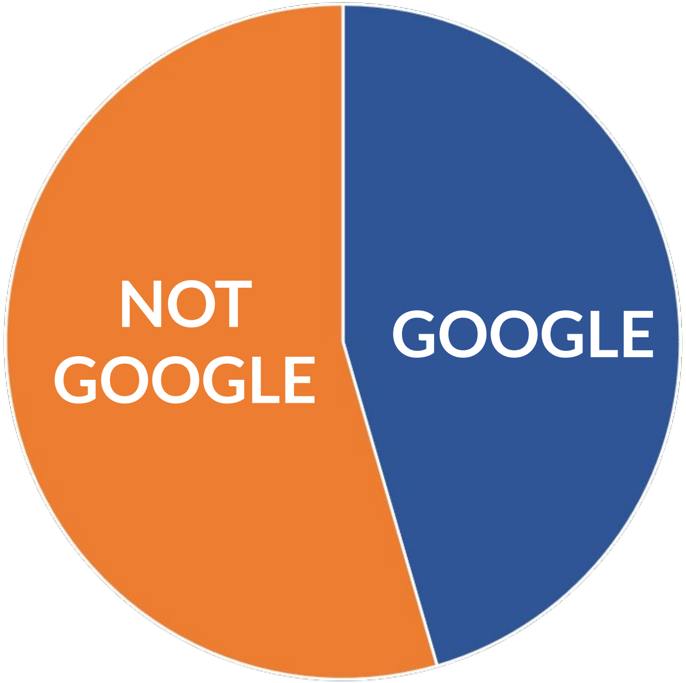




Community Contributions



Kubernetes



Kubeflow

Critical User Journey Comparison

2017

- Experiment with Jupyter
- Distribute your training with TFJob
- Serve your model with Seldon

2019

- Setup locally with miniKF
- Access your cluster with Istio/Ingress
- Transform your data with TF.T
- Analyze the data with TF.DV
- Experiment with Jupyter
- Hyperparam sweep with Katib
- Distribute your training with TFJob
- Analyze your model with TF.MA
- Serve your model with Seldon
- Orchestrate everything with KF.Pipelines

Just a SMALL sample of contributions

Arrikto

- Jupyter manager ui
- Pipelines volume support

Cisco

- Katib
- KubeBench
- PyTorch

GoJEK

- Feast feature store

IBM

- Pipeline components for spark, ffdl, Watson

Intel

- kfctl (CLI & library) & kustomize
- OpenVino

Intuit

- Argo

RedHat + NVIDIA

- TensorRT for notebooks

Seldon

- Seldon core

Introducing Kubeflow 0.5

What landed in 0.5?

Notebook Improvements

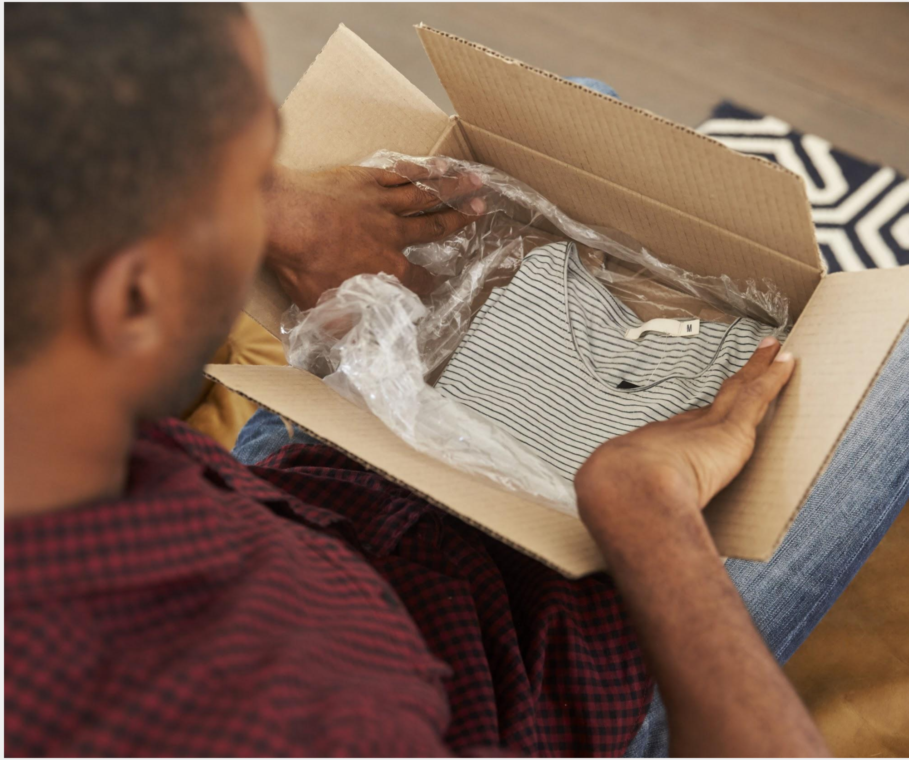
- New Jupyter UI & CR
- Multiple notebook support
- Build, train, deploy from notebook

Deployment

- Minikf for easy local install
- kfctl CLI and & go library

Pipelines

- GPU support
- Upgrade and external storage support
- TFX integration



Three 0.5 Features to Highlight

- Reducing the leap from exploration to production
- Notebook-based provisioning
- Kubeflow Pipelines integration

Demo

Dev to Prod with Kubeflow

- Prototype a model using a notebook
- Scale out using fairing
- Train and validate using pipelines that are built for production

Demo Video

Demo Recap

Dev to Prod with Kubeflow

Make data scientists happy

- Stay in a notebook
- Leverage K8s for scalability (batch jobs, scaling, etc...)

Make SRE happy

- Declarative, repeatable processes
- GitOps

Don’t rewrite notebook to deploy it

What’s coming in 0.6?

Enterprise readiness

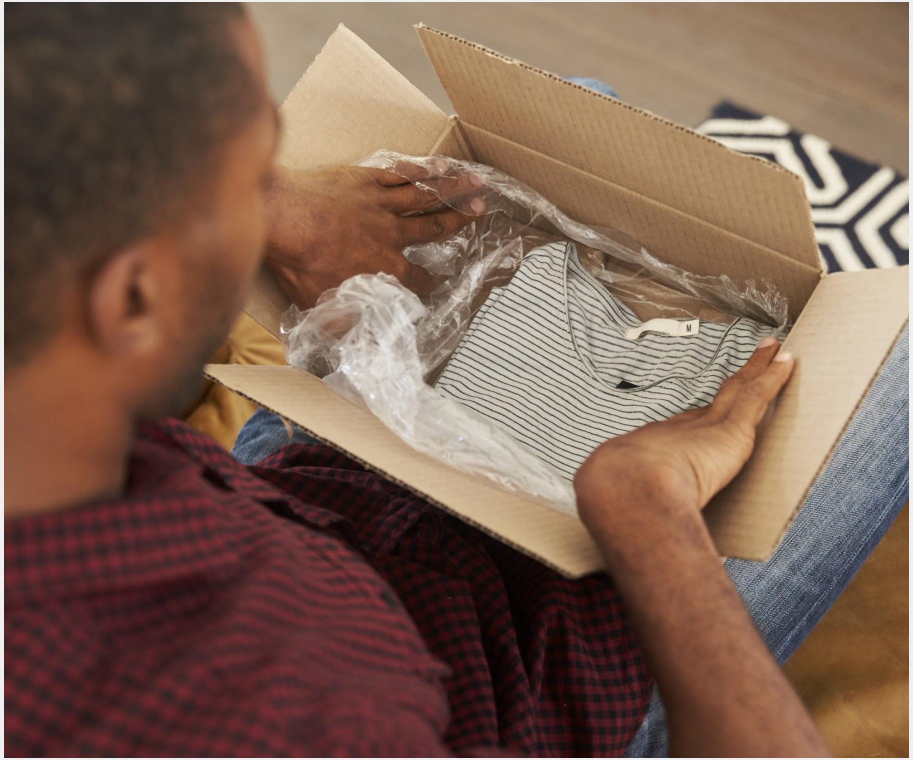
- Multi-user support
- ISTIO for service mesh and AuthZ
- API stability - TFJob & PyTorch 1.0

Advanced composability & tooling

- New metadata backend and UI for automated experiment tracking
- Replacing ksonnet with kustomize
- Katib - new UI, API and ML terminology

Pipelines

- Volume support
- Tensorboard management
- Metadata integration



Wasn’t This Supposed To Be A Talk about Kubeflow 1.0?

Good News!

ALREADY Production-Ready!

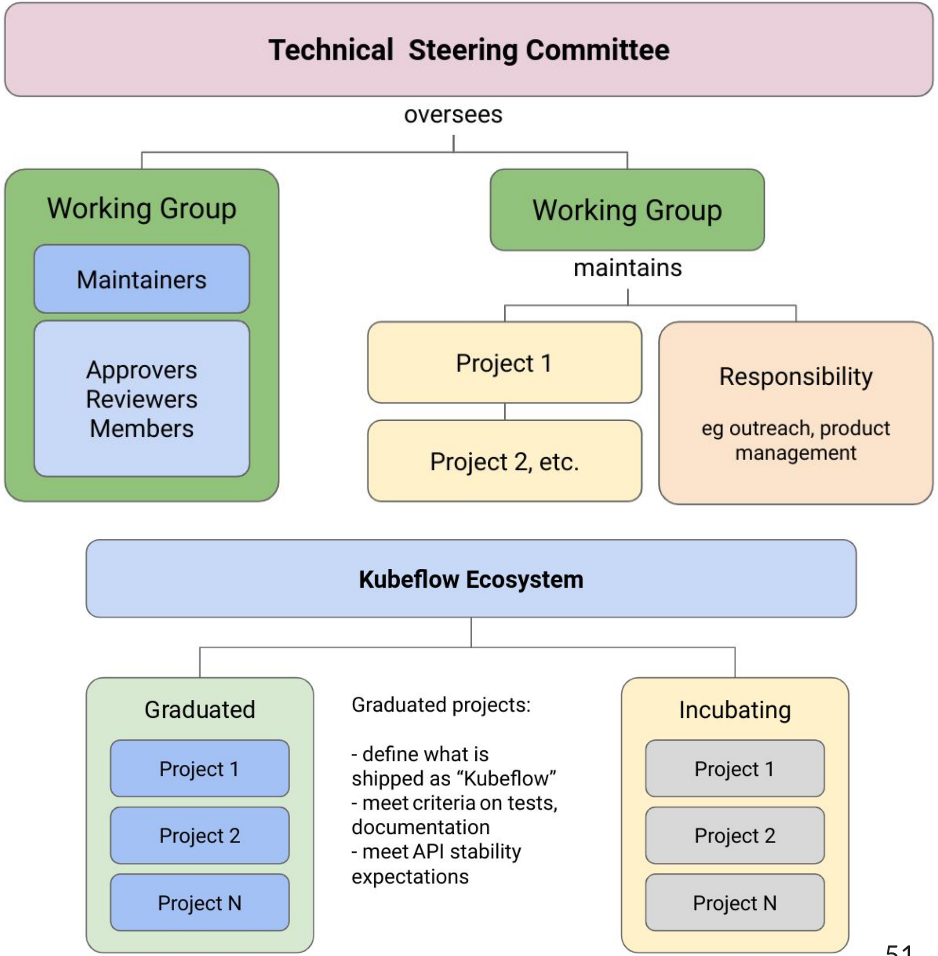
- Kubernetes
- TensorFlow & PyTorch
- TFX (TensorFlow Extended)
  - TensorFlow Transform
  - TensorFlow Data
  - TensorFlow Serving
- Ambassador/Istio
- Seldon

Being Thoughtful About 1.0

- We want to make sure we got the APIs correct to provide stability
- ALSO want to make sure we're nailing the critical user journeys
    - Build, train and deploy models from notebook
    - Multiple users/teams can share a Kubeflow cluster
    - Easy & uniform experience across multiple clouds
    - Rich pipelines for real MLOps
    - Artifact tracking and reproducibility
- For more info see full roadmap
    - https://github.com/kubeflow/kubeflow/blob/master/ROADMAP.md

Governance

- Ensure a sustainable and open community
- Refreshing governance
- http://bit.ly/kf_governance_proposal



Graduated projects:

- define what is shipped as "Kubeflow"
- meet criteria on tests, documentation
- meet API stability expectations

It’s a whole new world

- Data science will touch **EVERY** industry.
- We can’t ask people to become a PhD in statistics though.
- How do **WE** help everyone take advantage of this transformation?

Kubeflow is open!

Open comm-unity

Open design

Open to ideas

Open source

Come Help!

- website: https://kubeflow.org
- github: https://github.com/kubeflow/kubeflow
- slack: kubeflow (http://kubeflow.slack.com)
- twitter: @kubeflow

David Aronchick @aronchick
(david.aronchick@microsoft.com)

Jeremy Lewi (jlewi@google.com)

Kubeflow Talks (bit.ly/kf_calendar)

- **Tutorial Introduction to Pipelines** - *Tuesday May 21 14:00-15:25*; Michelle Casbon, Dan Sanche, Dan Anghel & Michal Zylinski Google (https://sched.co/MPgr)
- **Kubeflow BOF** - *Tuesday May 21 15:55-16:30*; David Aronchick, Microsoft & Yaron Haviv, Iguazio (https://sched.co/PiUF)
- **Toward Kubeflow 1.0, Bringing a Cloud Native Platform for ML to Kubernetes** - *Wednesday May 22 11:55 - 12:30*; David Aronchick, Microsoft & Jeremy Lewi Google (https://sched.co/MPax)
- **Building Cross-Cloud ML Pipelines with Kubeflow with Spark & TensorFlow** - *Wednesday May 22 14:00 - 14:35*; Holden Karau, Google & Trevor Grant, IBM (https://sched.co/MPaZ)
- **Managing Machine Learning Pipelines In Production with Kubeflow with Devops** - *Wednesday May 22 14:40-14:35* - David Aronchick, Microsoft (https://sched.co/MPaZ)
- **Large Scale Distributed Deep Learning with Kubernetes Operators -** *Wed May 22 15:55 - 16:30;* Yuan Tang, Ant Financial & Yong Tang MobileIron (https://sched.co/MPaT)
- **Moving People and Products with Machine Learning on Kubeflow** - Thursday May 23 14:00 -14:35; Jeremy Lewi, Google & Willem Pienaar, GO-JEK (https://sched.co/MPac)

Kubeflow

Thank You

www.kubeflow.org

github.com/jlewi/kubecon-demo